GenAttack: Practical Black-box Attacks with Gradient-Free Optimization

Project 5 - task 2

Moustafa Alzantot
UCLA
malzantot@ucla.edu

Yash Sharma
The Cooper Union
sharma2@cooper.edu

Supriyo Chakraborty
IBM Research
supriyo@us.ibm.com

Mani Srivastava
UCLA
mbs@ucla.edu

Abstract

Deep neural networks (DNNs) are vulnerable to adversarial examples, even in the black-box case, where the attacker is limited to solely query access. Existing black-box approaches to generating adversarial examples typically require a significant amount of queries, either for training a substitute network or estimating gradients from the output scores. We introduce `GenAttack`, a gradient-free optimization technique which uses genetic algorithms for synthesizing adversarial examples in the black-box setting. Our experiments on the MNIST, CIFAR-10, and ImageNet datasets show that `GenAttack` can successfully generate visually imperceptible adversarial examples against state-of-the-art image recognition models with orders of magnitude fewer queries than existing approaches. For example, in our CIFAR-10 experiments, `GenAttack` required roughly 2,568 times less queries than the current state-of-the-art black-box attack. Furthermore, we show that `GenAttack` can successfully attack both the state-of-the-art ImageNet defense, ensemble adversarial training, and non-differentiable, randomized input transformation defenses. `GenAttack`'s success against ensemble adversarial training demonstrates that its *query efficiency* enables it to exploit the defense's weakness to direct black-box attacks. `GenAttack`'s success against non-differentiable input transformations indicates that its *gradient-free nature* enables it to be applicable against defenses which perform gradient masking/obfuscation to confuse the attacker. Our results suggest that population-based optimization opens up a promising area of research into effective gradient-free black-box attacks.